Tecniche di deep learning per segnali audio

A cura di Valmacco Daniele

Matricola 816371

Università degli studi di Milano-Bicocca

Riconoscimento delle Emozioni nei Segnali Audio

- Ricerca di dataset
- Individuato features handcrafted per approcci di machine learning tradizionali
- Utilizzo di una rete deep pretrainata per generare delle features
- Analisi di diversi classificatori utilizzando features handcrafted e features deep.

Introduzione agli Studi Relativi al Dataset

- Il dataset deve comprendere un numero di emozioni non eccessivo, ma nemmeno troppo poco numeroso.
- Queste devono essere inoltre non fraintendibili e devono coprire la totalità dello spettro emotivo che può essere indotto durante l'ascolto di un brano musicale.
- La scala migliore, nonché anche la più utilizzata al momento, è la scala GEMS*: consiste nel raggruppamento in nove categorie dei principali stati emotivi raccolti.

1st Order Factors	Items
Wonder	Moved
	Filled with Wonder
	Allured
Trascendence	Fascinated
	Overwelmed
	Feeling of trascendence
Peacefulness	Serene
	Calm
	Soothed
Tenderness	Tender
	Affectionate
	Mellow
Nostalgia	Nostalgic
	Sentimental
	Dreamy
Power	Strong
	Energetic
	Triumphant
Joyful Activation	Animated
	Bouncy
	Joyful
Sadness	Sad
	Tearful
	Blue
Tension	Tense
	Agitated
	Nervous

* M. Zentner, D. Grandjean, e K. R. Scherer, «Emotions evoked by the sound of music: Characterization, classification, and measurement.», Emotion, vol. 8, n. 4, pagg. 494–521, 2008, doi: 10.1037/1528-3542.8.4.494.

Dataset Emotify

- Emotify* è un ‘Game With A Purpose’ in cui gli utenti sono invitati ad ascoltare una traccia musicale della lunghezza di un minuto, al termine della quale sono invitati ad inserire al più tre tra le emozioni mostrate.

- Sono disponibili sia i file musicali utilizzati (400 brani, divisi per genere musicale), che il ground truth.

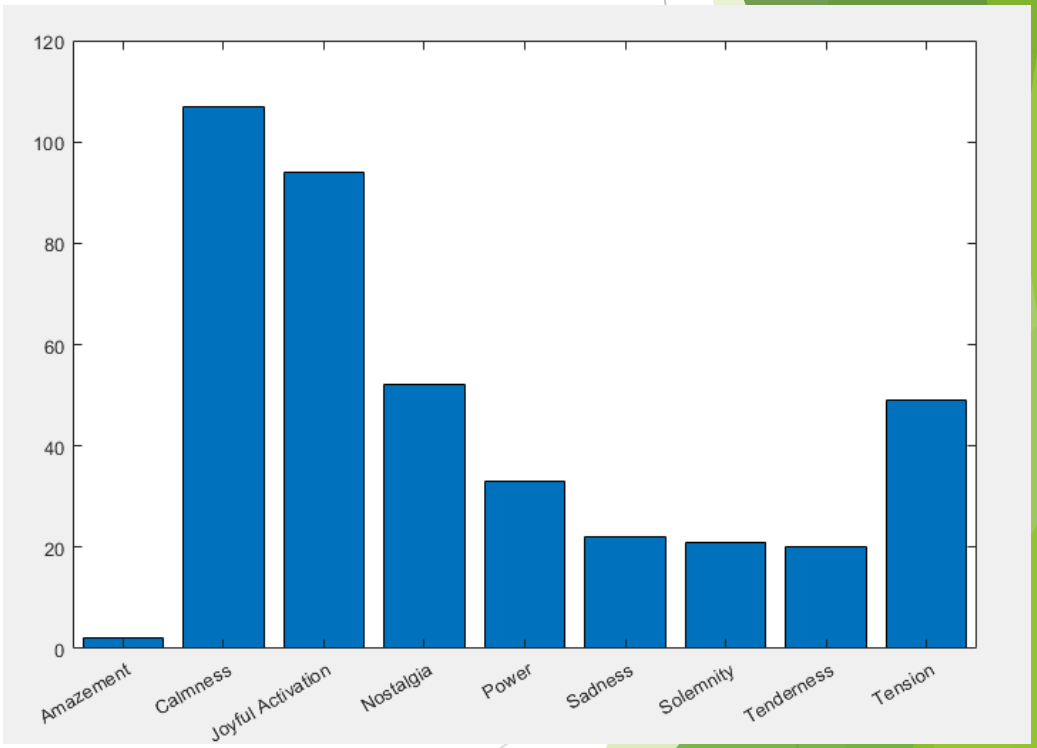
* A. Aljanaki, F. Wiering, e R. C. Veltkamp, «Studying emotion induced by music through a crowdsourcing game», Inf. Process. Manag., vol. 52, n. 1, pagg. 115–128, gen. 2016, doi: 10.1016/j.ipm.2015.03.004.

Elaborazione dei Dati

- I dati estratti mediante Emotify sono stati poi elaborati in modo tale da avere una sola emozione dominante per brano.

- Per ovviare alla scarsa numerosità dei brani utilizzanti durante il gioco, è stato necessario un processo di Data Augmentation (ottenendo così 1588 istanze della durata di 15 secondi).

- La ridotta numerosità della classe ‘Amazement’ (2 brani) ha richiesto che questa venisse cancellata in un secondo momento (arrivando così a 1580 istanze).



Features Handcrafted

- È stato utilizzato MIRtoolbox*, un'estensione di MATLAB per estrarre features dai segnali audio.

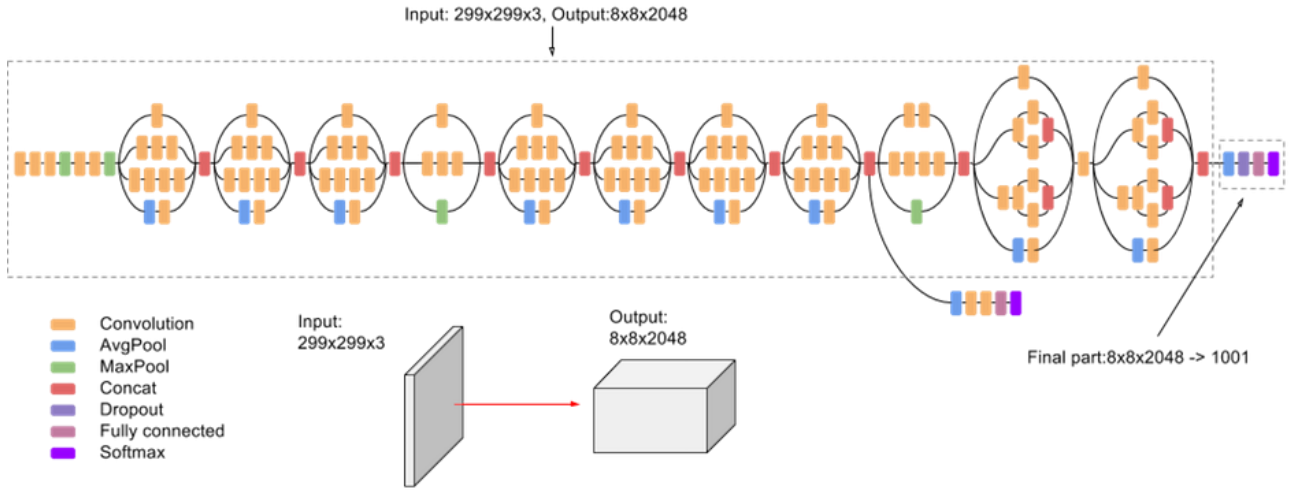
- Converte i file audio in oggetti 'miraudio' che verranno poi riconvertiti nei loro valori numerici sulla base delle features estratte.

- Durante questo processo sono state estratte 17 features riguardanti il segnale audio (come la RMS, Rolloff, la Chromagram, etc..).

* https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox/manual1-7.pdf

Features Deep (Inception-v3)

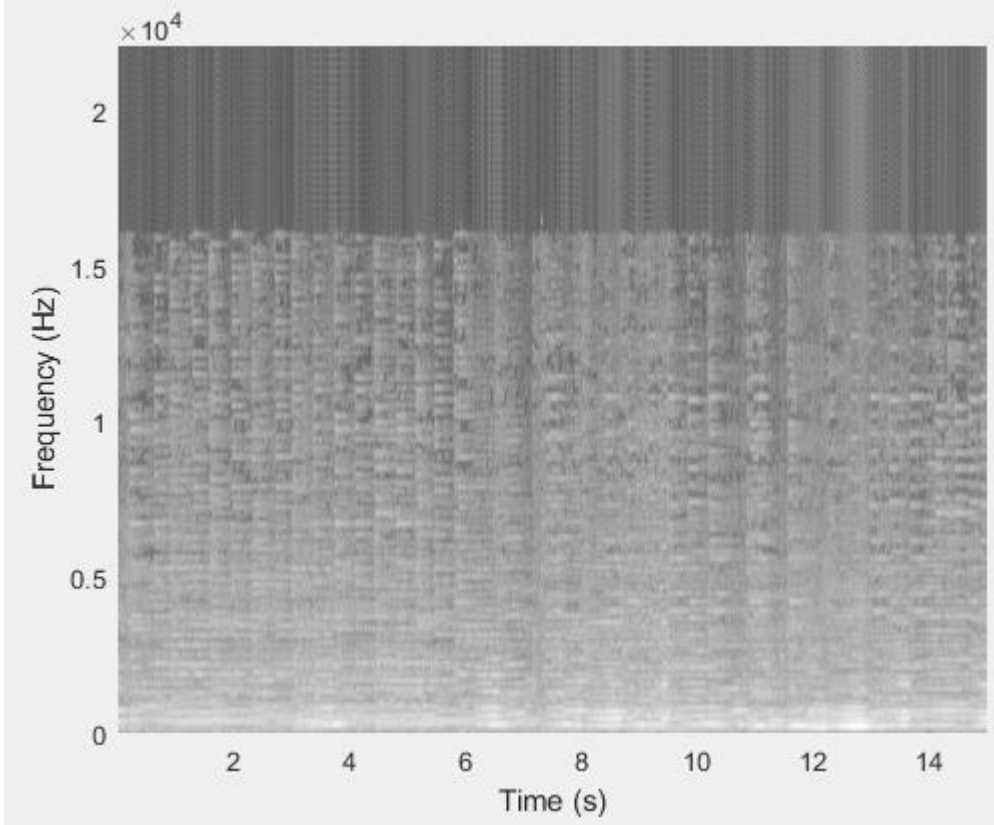
- Inception-v3* è una rete neurale convoluzionale pretrainata sul database ImageNet. In questo lavoro è stata utilizzata per estrarre features dagli spettrogrammi.
- Questa rete chiede in input immagini.
- Dalla rete neurale sono state estratte 2048 features da uno degli ultimi layers.



* C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, e Z. Wojna, «Rethinking the Inception Architecture for Computer Vision»

Spettrogrammi

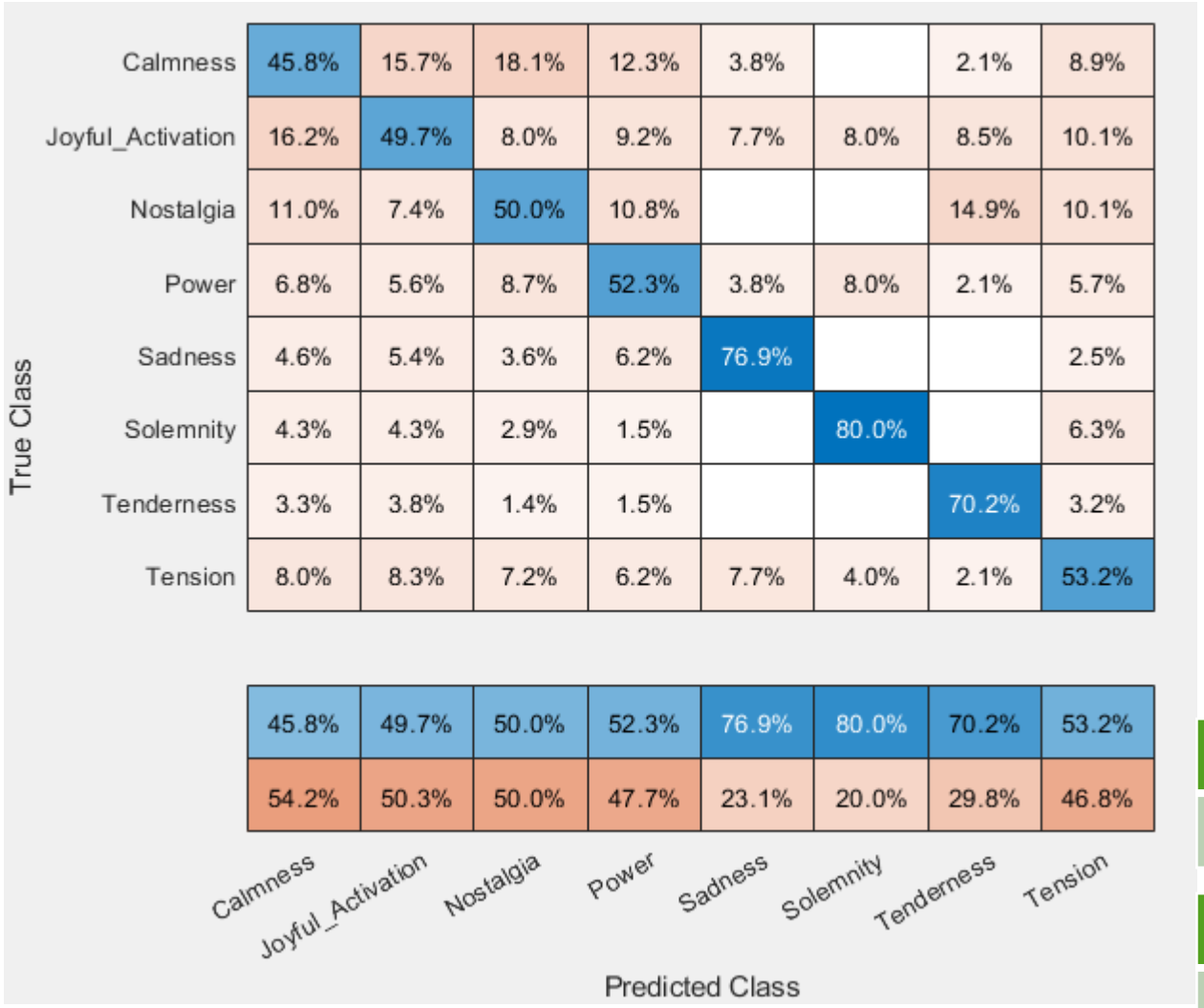
- È stato possibile ricavare gli spettrogrammi dalle tracce musicali.
- Il segnale è stato poi riscalato per adattarsi al range delle immagini (0 – 255) applicando una trasformazione logaritmica.
- Lo spettrogramma è stato ricampionato sia sull'asse x che sull'asse y per adattarsi alle dimensioni di input.
- L'immagine dello spettrogramma a livelli di grigio è stata replicata sui tre canali RGB.



Classificazione delle Emozioni

- Sono stati provati diversi approcci di classificazione (tra cui SVM, KNN ed Ensemble) sia con features handcrafted che con le features deep.
- Come metodo di validazione è stato scelto il k-fold cross-validation (k = 5).
- I dati sono stati precedentemente divisi in cinque fold in modo tale da non avere, in fase di test, alcune istanze della stessa canzone usate durante la fase di training.
- Le performance dei vari algoritmi sono state analizzate per mezzo dei valori di accuracy e tramite le matrici di confusione.

Features Handcrafted

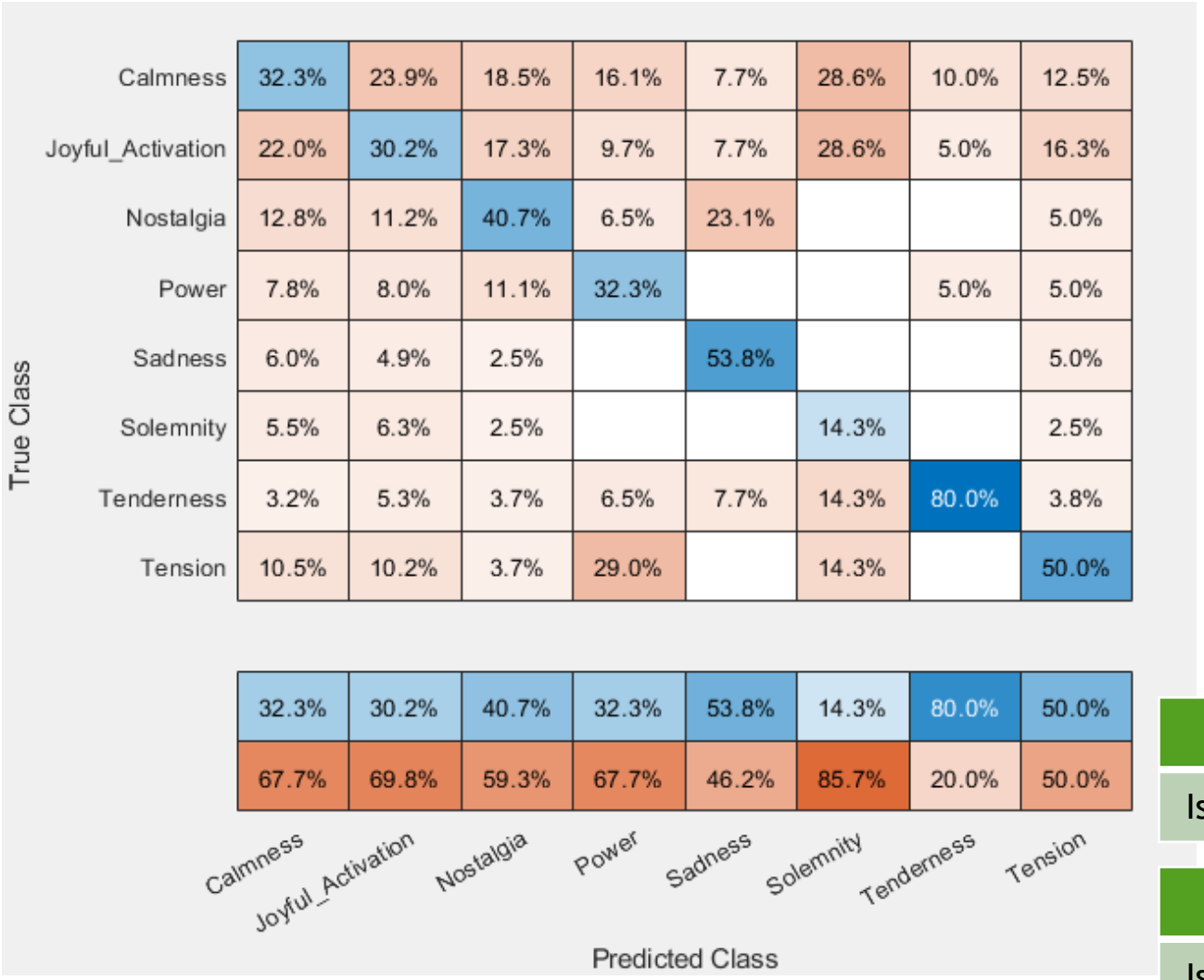


- Ensemble Bagged Tree
- K-fold = 5
- Accuracy = 50.1%
- Random Guess = 12.5%
- Features = 17

Class:	Calmness	Joyful A.	Nostalgia	Power
Istances:	444	368	174	122

Class:	Sadness	Solemnity	Tenderness	Tension
Istances:	96	63	68	245

Features Deep



- Cubic SVM
- K-fold = 5
- Accuracy = 33.7%
- Random Guess = 12.5%
- Features = 2048

Class:	Calmness	Joyful A.	Nostalgia	Power
Istances:	444	368	174	122

Class:	Sadness	Solemnity	Tenderness	Tension
Istances:	96	63	68	245

Conclusioni

- Le performance possono essere peggiorate in seguito a fattori come:
  - la disparità tra le classi
  - la soggettività con cui sono stati raccolti i dati
  - la mancanza di features relative al testo del brano
  - la possibilità di aver danneggiato il dataset durante la fase di data augmentation
- Per risolvere i problemi avrei potuto:
  - far valutare ogni istanza della canzone originale ottenuta tramite data augmentation
  - utilizzare dataset di dimensioni maggiori
  - apportare delle modifiche alla scala GEMS (es. utilizzando un numero minore di categorie)
  - utilizzare altri approcci che avrebbero potuto dare come risultato una distribuzione di classi anziché una classe sola